Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Tests for scale changes based on pairwise differences

Carina Gerstenberger, Daniel Vogel & Martin Wendler

To cite this article: Carina Gerstenberger, Daniel Vogel & Martin Wendler (2019): Tests for scale changes based on pairwise differences, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1629938

To link to this article: https://doi.org/10.1080/01621459.2019.1629938

View supplementary material

Accepted author version posted online: 19 Jun 2019.

Submit your article to this journal

Article views: 51

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

Check for updates

Tests for scale changes based on pairwise differences

Carina Gerstenberger*, Daniel Vogel†, and Martin Wendler‡
*Fakultät für Mathematik, Ruhr-Universität Bochum, 44801 Bochum, Germany
†Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom
‡Institut für Mathematik und Informatik, Universität Greifswald, 17489 Greifswald, Germany

Abstract

In many applications it is important to know whether the amount of fluctuation in a series of observations changes over time. In this article, we investigate different tests for detecting changes in the scale of mean-stationary time series. The classical approach, based on the CUSUM test applied to the squared centered observations, is very vulnerable to outliers and impractical for heavy-tailed data, which leads us to contemplate test statistics based on alternative, less outlier-sensitive scale estimators. It turns out that the tests based on Gini's mean difference (the average of all pairwise distances) and generalized Qn estimators (sample quantiles of all pairwise distances) are very suitable candidates. They improve upon the classical test not only under heavy tails or in the presence of outliers, but also under normality.

We use recent results on the process convergence of U-statistics and U-quantiles for dependent sequences to derive the limiting distribution of the test statistics and propose estimators for the long-run variance. We show the consistency of the tests and demonstrate the applicability of the new change-point detection methods at two real-life data examples from hydrology and finance.

Accepted Manuscript

Keywords: Block bootstrap, Gini's mean difference, Long-run variance estimation, U-quantile, U-statistic

1 Introduction

The established approach to testing for scale changes of a univariate time series $X_1,\ldots,X_n$ is a CUSUM test applied to the squares of the centered observations. The test statistic may be written as

$$\hat{T}_{\sigma^2} = \max_{1\le k\le n} \frac{k}{\sqrt{n}} \left|\hat{\sigma}^2_{1:k} - \hat{\sigma}^2_{1:n}\right|, \qquad (1)$$

where $\hat{\sigma}^2_{i:j}$ denotes the sample variance computed from the observations $X_i,\ldots,X_j, 1\le i<j\le n$. To carry out the test in practice, the test statistic is usually divided by (the square root of) a suitable estimator of the corresponding long-run variance. This has first been considered by Inclan and Tiao (1994), who derive asymptotics for centered, normal, i.i.d. data. It has subsequently been extended by several authors to broader situations, e.g., Gombay et al. (1996) allow the mean to be unknown and propose a weighted version of the testing procedure, Lee and Park (2001) extend it to linear processes, and Wied et al. (2012) extend it further to broader short-range dependence condition. A multivariate version was considered by Aue et al. (2009).

The test statistic (1) is prone to outliers. This has already been remarked by Inclan and Tiao (1994) and has led Lee and Park (2001) to consider a version of the test using trimmed observations. Outliers may affect the test decision in both directions: A single outlier suffices to make the test reject the null hypothesis at an otherwise stationary sequence, but more often one finds that outliers mask a change, and the test is generally very inefficient at heavy-tailed population distributions.

Writing the test statistic as in (1) suggests that this behavior may be largely attributed to the use of the sample variance as a scale estimator. The recognition of the extreme "non-robustness" of the sample variance and

Accepted Manuscript

derived methods, in fact, stood at the beginning of the development of the area of robust statistics as a whole (e.g. Tukey, 1960). Thus, an intuitive way of constructing robust scale change-point tests is to replace the sample variance by an alternative scale estimator.

Before surveying potential alternative scale measures, we introduce some general concepts and notation. Le $\mathcal{L}(X)$ to denote the *law*, i.e., the distribution, of any random variable *X*. We call any function $s:\mathcal{F}\to[0,\infty]$, where $\mathcal{F}$ is the set of all univariate distributions *F*, a *scale measure* (or *dispersion measure*) if it satisfies $s(\mathcal{L}(aX+b))=|a|s(\mathcal{L}(X))$ for all $a,b\in$ R. Although not being a scale measure itself, the variance $\sigma^2=E(X-EX)^2$ is, in a lax use of the term, often referred to as such, since it is closely related to the standard deviation.

A *scale estimator* s_n is then generally understood as the sample version $s(F_n)$ of $s(F)$, where F_n is the empirical distribution associated with the data set $X_1,\ldots,X_n$. Letting s_n be a scale estimator and s the corresponding population value, the asymptotic variance $ASV(s_n;F)$ of s_n at the distribution F is defined as the variance of the limiting normal distribution of $\sqrt{n}(s_n-s)$, when s_n is evaluated at an independent sample $X_1,\ldots,X_n$ drawn from *F*. In order to make two scale estimators $s_{1,n}$ and $s_{2,n}$ comparable efficiency-wise, we have to normalize them appropriately, and define their asymptotic relative efficiency at the population distribution *F* as $ARE(s_{1,n},s_{2,n};F)=ASV(s_{2,n};F)/ASV(s_{1,n};F)(s_1/s_2)^2$, where s_1 and s_2 denote the respective population values at *F*.

Let $\mathrm{md}(F)$ denote the median of the distribution *F*, i.e., the center point of the interval $\{x\in \mathrm{R}\mid F(x-)\le 1/2\le F(x)\}$, where $F(x-)$ denotes the left-hand limit. We define the *mean deviation* as $d(F)=E|X-\mathrm{md}(F)|$ and its empirical version as $d_n=\frac{1}{n-1}\sum_{i=1}^{n}|X_i-\mathrm{md}(F_n)|$. Its asymptotic relative efficiency with respect to the standard deviation is 88% at normality. However, Tukey (1960)

Accepted Manuscript

pointed out that it is more efficient than the standard deviation if the normal distribution is slightly contaminated. So, the mean deviation may be a suitable candidate for constructing less outlier-sensitive change-point tests. Gerstenberger and Vogel (2015) argue that, when pondering the mean deviation instead of the standard deviation for robustness reasons, it may be better to use Gini's mean difference $g_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|$, i.e., the mean of the absolute distances of all pairs of observations. The population version is $g(F) = E|X - Y|$, where $X, Y \sim F$ are independent. Gini's mean difference exhibits a comparable sensitivity towards heavy tails as the mean deviation, but retains an asymptotic relative efficiency with respect to the standard deviation of 98% at the normal distribution (Nair, 1936).

Both estimators, the mean deviation and Gini's mean difference, improve upon the standard deviation in terms of robustness, but are not robust in a modern understanding of the term. Both have, e.g., an unbounded influence function (Gerstenberger and Vogel, 2015). A highly robust scale estimator is the median absolute deviation (MAD), popularized by Hampel (1974). The population value $m(F)$ is the median of the distribution of $|X - \mathrm{md}(F)|$ and the sample version $m_n = m_n(\mathrm{X}_n)$ is the median of the values $|X_i - \mathrm{md}(F_n)|, 1 \le i \le n$. The MAD has a bounded influence function (see Huber and Ronchetti, 2009, Section 6.4) and a breakdown point of about 50%. Its main drawback is its poor asymptotic efficiency under normality, which is only 37% as compared to the standard deviation. It is also unsuitable for change-in-scale detection due to other reasons that will be detailed in Sections 2.3 and 5. (For the robustness concepts *influence function* and *breakdown point*, see, e.g., Huber and Ronchetti (2009), Sections 1.5 and 11.2, respectively, or Maronna et al. (2006), Chapter 3.)

Similarly to going from the *mean* deviation to the *median* absolute deviation, we may consider the median, or more generally any sample α-quantile, of all pairwise differences. We call this estimator Q_n^α and the corresponding population scale measure Q^α, i.e., $Q^\alpha = U^{-1}(\alpha) = \inf\{x | \alpha \le U(x)\}$, where U is

Accepted Manuscript

the distribution function of $|X-Y|$ for $X,Y \sim F$ independent, and U^{-1} the corresponding quantile function. For the precise definition of Q_n^α, any sensible definition of the sample quantile can be employed, see, e.g., Hyndman and Fan (1996). The asymptotic results we derive later are not affected by this choice, and any practical differences turn out to be negligible. For simplicity, we define $Q_n^\alpha = U_n^{-1}(\alpha)$, where U_n is the empirical distribution function associated with the sample $|X_i - X_j|, 1 \le i < j \le n$. In case of Gini's mean difference, we have observed that the transition from the average distance from the symmetry center to the average pairwise distance led to an increase in efficiency under normality. The effect is even more pronounced for the median distances: we have $ARE(Q_n^{0.5}, \sigma_n, N(0,1)) = 86.3\%$. Rousseeuw and Croux (1993) propose to use the lower quartile, i.e., $\alpha = 1/4$, instead of the median. They call this estimator Q_n, and it has become known under this name, which leads us to call the generalized version Q_n^α. Rousseeuw and Croux's choice of $\alpha = 1/4$ is motivated by high-breakdown-point considerations. This aspect is of much lesser relevance for the change-point problem, for which the original Q_n is indeed unsuitable. A larger α-value of roughly $0.7 < \alpha < 0.9$ is much more appropriate. We defer further explanations to Section 2.3.

These five scale measures, the standard deviation σ_n, the mean deviation d_n, Gini's mean difference g_n, the median absolute deviation (MAD) m_n, and the α-sample quantile of all pairwise differences Q_n^α, are the ones we restrict our attention to in the present article. There are many more potential scale estimators that satisfy the above scale equivariance and many more proposals in the robustness literature, many of which require a data-adaptive re-weighting of the observations, see, e.g., Huber and Ronchetti (2009, Chapter 5) or Lax (1985). We explore the use of these common, easy-to-compute estimators for change-point testing. They all admit explicit formulas, all can be computed in $O(n \log n)$ time, and the pairwise-difference estimators allow computing time savings for sequentially updated estimates (which are required in the change-point setting) – more so than, e.g., implicitly defined

Accepted Manuscript

estimators. The two pairwise-difference based estimators, the average and the α-quantile of all pairwise differences, possess promising statistical properties. They are almost as efficient as the standard deviation at normality and, hence, the improvement in robustness is expected to come at practically no loss in terms of power under normality. In fact, as it turns out, these tests can have a better power than the variance-based test also under normality.

The paper is organized as follows: Section 2 states the test statistics and long-run variance estimators with Subsection 2.3 discussing the choice of α for the Q_n^α. Section 3 contains theoretical results: we study the asymptotic behavior of the pairwise-differences-based test statistics under stationarity as well as under fixed-size alternatives. Section 4 addresses practical aspects: a rule for data-adaptive bandwidth selection for the long-run variance estimation is outlined, and, as an alternative to studentization, a block bootstrap scheme is described. Both are implemented in the simulations presented in Section 5. Section 6 illustrates the behavior of the tests at real-life data examples, and Section 7 concludes. The paper is accompanied by an online supplement, which contains background information on the short-range condition *P*-NED (Section A), the derivations for the asymptotic results for the Q_n^α and other scale estimators used in Section 2.3 (Section B), and the proofs for Section 3 (Section C).

2 Test statistics and long-run variance estimates

We first describe the data model employed (Section 2.1). We then propose several change-point test statistics based on the scale estimators introduced and provide estimates for their long-run variances (Section 2.2). Finally, we discuss the selection of α for the Q_n^α (Section 2.3).

2.1 The data model

We assume the data $X_1,\ldots,X_n$ to follow the model $X_i = \lambda_i Y_i + \mu, 1 \le i \le n$, where $Y_1,\ldots,Y_n$ are part of the stationary, median-centered sequence $(Y_i)_{i\in Z}$. We want to test the hypothesis $H_0 : \lambda_1 = \ldots = \lambda_n$ against the alternative

Accepted Manuscript

$H_1 : \exists k \in \{1,\ldots,n-1\} : \lambda_1 = \ldots = \lambda_k \neq \lambda_{k+1} = \ldots = \lambda_n$. This set-up is completely moment-free. We allow the underlying process to be dependent, more precisely we assume $(Y_i)_{i\in\mathbb{Z}}$ to be near-epoch dependent in probability on an absolutely regular process. We briefly introduce this short-range dependence condition.

Definition 2.1. *Let $\mathcal{A},\mathcal{B} \subset \mathcal{F}$ be two σ-fields on the probability space $(\Omega,\mathcal{F},\mathrm{P})$. The absolute regularity coefficient of $\mathcal{A}$ and $\mathcal{B}$ is $\beta(\mathcal{A},\mathcal{B}) = E(\operatorname{esssup}_{A\in\mathcal{A}} |\mathrm{P}(A\mid\mathcal{B}) - \mathrm{P}(A)|)$. For a stationary process $(Z_n)_{n\in\mathbb{Z}}$, the absolute regularity coefficients of $(Z_n)_{n\in\mathbb{Z}}$ are given by $\beta_k = \sup_{n\in\mathbb{N}} \beta(\sigma(Z_1,\ldots,Z_n), \sigma(Z_{n+k}, Z_{n+k+1},\ldots))$. We say that $(Z_n)_{n\in\mathbb{Z}}$ is absolutely regular, if $\beta_k \to 0$ as $k\to\infty$.*

Mixing assumptions such as absolute regularity are difficult to check in practice and do not necessarily hold even for simple models like an autoregressive process. So we will not study absolutely regular processes themselves, but approximating functionals of such processes. In this situation, L_2 near-epoch dependence is frequently used. However, since we also consider quantile-based estimators with the advantage of moment-freeness, we want to avoid any moment assumptions implicitly hidden in the short-range conditions. For this reason, we employ the concept of *near-epoch dependence in probability*, introduced by Dehling et al. (2017). For further information see Appendix A in the supplementary material.

Definition 2.2. *We call the process $(X_n)_{n\in\mathbb{N}}$ near-epoch dependent in probability (or short P-NED) on the process $(Z_n)_{n\in\mathbb{Z}}$ if there is a sequence of approximating constants $(a_l)_{l\in\mathbb{N}}$ with $a_l \to 0$ as $l\to\infty$, a sequence of functions $f_l : \mathbb{R}^{2l+1} \to \mathbb{R}$ and a non-increasing function $\Phi : (0,\infty) \to (0,\infty)$ such that*

$$\mathrm{P}\left(\left|X_0 - f_l(Z_{-l},\ldots,Z_l)\right| > \epsilon\right) \leq a_l \Phi(\epsilon) \quad (2)$$

Accepted Manuscript

for all $l \in \mathbb{N}$ *and* $\epsilon > 0$.

The absolute regularity coefficients β_k and the approximating constants a_l will have to fulfill certain rate conditions that are detailed in Assumptions 3.1 and 3.2.

2.2 Change-point test statistics and long-run variance estimates

We test the null hypothesis H_0 against the alternative H_1 by means of CUSUM-type test statistics of the form $\hat{T}_s = \max_{1\le k\le n} \frac{k}{\sqrt{n}} \left| s_{1:k} - s_{1:n} \right|$. Throughout, we use s_n as generic notation for a scale estimator (where we include the variance), and $s_{1:k}$ denotes the estimator applied to $X_1, \ldots, X_k$. For the scale estimators introduced in Section 1, we obtain the test statistics $\hat{T}_{\sigma^2}, \hat{T}_d, \hat{T}_g, \hat{T}_m$, and $\hat{T}_Q(\alpha)$, respectively. Under the null hypothesis H_0, the sequence $X_1, \ldots, X_n$ is stationary, and can be thought of as being part of a stationary process $(X_i)_{i\in\mathbb{Z}}$ with marginal distribution F. Then, under suitable regularity conditions (that are specific to the choice of s_n), the test statistic $\hat{T}_s$ converges in distribution to $D_s \sup_{0\le t\le 1} |B(t)|$ as $n \to \infty$, where B is a Brownian bridge. The quantity D_s^2 is referred to as the long-run variance. Expressions for the scale estimators considered here are given below. The distribution of $\sup_{0\le t\le 1} |B(t)|$ is well known and referred to as Kolmogorov distribution. However, D_s^2 is generally unknown, depends on the distribution of the whole process $(X_i)_{i\in\mathbb{Z}}$ and must be estimated when applying the test in practice. Alternatively, bootstrapping can be employed. This is discussed in Section 4.2.

In the following definitions, let $X, Y \sim F$ be independent. The long-run variances corresponding to the scale estimators under consideration are

$$D_{\sigma^2}^2 = \sum_{h=-\infty}^{\infty} \operatorname{cov}\left\{(X_0 - EX_0)^2, (X_h - EX_h)^2\right\},$$
$$D_d^2 = \sum_{h=-\infty}^{\infty} \operatorname{cov}\left(|X_0 - \operatorname{md}(F)|, |X_h - \operatorname{md}(F)|\right), \quad D_g^2 = 4\sum_{h=-\infty}^{\infty} \operatorname{cov}\left(\varphi(X_0), \varphi(X_h)\right), \tag{3}$$

Accepted Manuscript

where $\varphi(x) = E|x - Y| - g(F)$,

$$D_m^2 = \frac{1}{f_Z^2(m(F))} \sum_{h=-\infty}^{\infty} \operatorname{cov}\left(1_{\{|X_0 - \mathrm{md}(F)| \le m(F)\}},\ 1_{\{|X_h - \mathrm{md}(F)| \le m(F)\}}\right),$$

where f_Z is the density of $Z = |X - \mathrm{md}(F)|$, and

$$D_Q^2(\alpha) = \frac{4}{u^2(Q^\alpha(F))} \sum_{h=-\infty}^{\infty} \operatorname{cov}\left(\psi(X_0), \psi(X_h)\right), \qquad (4)$$

where $\psi(x) = P(|x - Y| \le Q^\alpha) - \alpha$ and $u(t)$ is the density associated with the distribution function $U(t) = P(|X - Y| \le t)$ of $|X - Y|$. An intuitive derivation of the expressions for D_g^2 and $D_Q^2(\alpha)$ are given in Appendix C in the supplementary material.

The following long-run variance estimators follow the construction principle of heteroscedasticity and autocorrelation consistent (HAC) kernel estimators, for which we use results by de Jong and Davidson (2000). The HAC kernel function (or weight function) W can be quite general, but has to fulfill Assumption 2.1 (a) below, which is basically Assumption 1 of de Jong and Davidson (2000). There is further a bandwidth to choose, which has to fulfill the rate condition of Assumption 2.1 (b) for the long-run variance estimator to be consistent.

Assumption 2.1.

(a) The function $W : [0, \infty) \to [-1, 1]$ *is continuous at 0 and at all but a finite number of points and* $W(0) = 1$. *Furthermore,* $|W|$ *is dominated by a non-increasing, integrable function and* $\int_0^\infty \left| \int_0^\infty W(t) \cos(xt)\, dt \right| dx < \infty$.

(b) The bandwidth b_n *satisfies* $b_n \to \infty$ *as* $n \to \infty$ *and* $b_n / \sqrt{n} \to 0$.

We propose the following long-run variance estimators for the three moment-based scale measures: For the variance we take a weighted sum of empirical autocorrelations of the centered squares of the data, i.e.,

Accepted Manuscript

$$\hat{D}^2_{\sigma^2} = \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \left\{(X_i - \bar{X}_n)^2 - \sigma_n^2\right\} \left\{(X_{i+|k|} - \bar{X}_n)^2 - \sigma_n^2\right\}, \quad (5)$$

where $\bar{X}_n$ and σ_n^2 denote the sample mean and the sample variance, respectively, computed from the whole sample. Similar expressions have been considered, e.g., by Gombay et al. (1996), Lee and Park (2001) and Wied et al. (2012). For the mean deviation, we propose

$$\hat{D}^2_d = \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \left(|X_i - \mathrm{md}(F_n)| - d_n\right)\left(|X_{i+|k|} - \mathrm{md}(F_n)| - d_n\right),$$

where $md(F_n)$ and d_n denote the sample median and the sample mean deviation, respectively, of the whole sample. For Gini's mean difference, we consider

$$\hat{D}^2_g = 4 \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \hat{\varphi}_n(X_i)\hat{\varphi}_n(X_{i+|k|}),$$

where $\hat{\varphi}_n(x) = n^{-1}\sum_{i=1}^{n}|x - X_i| - g_n$ is an empirical version of $\varphi(x)$ in (3). For the long-run variance estimates for the quantile-based scale measures m_n and Q_n^α, we need estimates for the densities f_Z and u, respectively, for which we use kernel density estimates

$$\hat{f}_Z(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{|X_i - \mathrm{md}(F_n)| - t}{h_n}\right), \quad \hat{u}(t) = \frac{1}{\binom{n}{2} h_n} \sum_{1 \le i < j \le n} K\left(\frac{|X_i - X_j| - t}{h_n}\right).$$

The density kernel K and the bandwidth h_n have to fulfill the following conditions.

Assumption 2.2. *Let* $K : \mathrm{R} \to \mathrm{R}$ *be symmetric (i.e.* $K(x) = K(-x)$*), Lipschitz-continuous function with bounded support which is of bounded variation and integrates to 1. Let the bandwidth* h_n *satisfy* $h_n \to 0$ *and* $nh_n^{8/3} \to \infty$, *as* $n \to \infty$.

Accepted Manuscript

Letting $\hat{\xi}_n(x) = 1_{\{|x-\mathrm{md}(F_n)|\le m_n\}} - 1/2$ and $\hat{\psi}_n(x) = n^{-1}\sum_{i=1}^{n} 1_{\{|x-X_i|\le Q_n^{\alpha}\}} - \alpha$, we define

$$\hat{D}_m^2 = \frac{1}{\hat{f}_Z(m_n)} \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \hat{\xi}_n(X_i)\hat{\xi}_n(X_{i+|k|})$$

and

$$\hat{D}_Q^2(\alpha) = \frac{4}{\hat{u}(Q_n^{\alpha})} \sum_{k=-(n-1)}^{n-1} W\left(\frac{|k|}{b_n}\right) \frac{1}{n} \sum_{i=1}^{n-|k|} \hat{\psi}_n(X_i)\hat{\psi}_n(X_{i+|k|}).$$

In Section 3, we give sufficient conditions for the convergence of the studentized test statistics $\hat{D}_g^{-1}\hat{T}_g$ and $\hat{D}_Q^{-1}(\alpha)\hat{T}_Q(\alpha)$, respectively, since the corresponding estimators, as outlined in Section 1, exhibit the best statistical properties, and these tests indeed show the best performance, as demonstrated in Section 5. The variance-based test statistic $\hat{D}_{\sigma^2}^{-1}\hat{T}_{\sigma^2}$, or versions of it, has been considered by several authors, e.g., it is treated for L_2 NED on α-mixing processes by Wied et al. (2012). As for the mean-deviation-based test statistic $\hat{D}_d^{-1}\hat{T}_d$, the convergence can be shown by similar techniques as for $\hat{D}_{\sigma^2}^{-1}\hat{T}_{\sigma^2}$: the same $(2+\delta)$ moment condition as for Gini's mean difference along with corresponding rate for the short-range dependence conditions (Assumption 3.1) are required. Additionally, a smoothness condition around $\mathrm{md}(F)$ is necessary to account for the estimation of the central location. For the MAD-based test statistic $\hat{D}_m^{-1}\hat{T}_m$, no moment conditions are required, but smoothness conditions on F at $\mathrm{md}(F)$ as well as $m(F) = |X - \mathrm{md}(F)|, X \sim F$. However, it turns out that the MAD does not provide a workable change-point test. Roughly speaking, the estimate is rather coarse, and the convergence to the limit distribution is too slow to yield usable critical values. But even for large n or with the use of bootstrapping methods, the test is dominated in terms of power by the other tests considered.

2.3 The choice of α for Q_n^{α}

Accepted Manuscript

We address the question which α to choose when employing the Q_n^α scale estimator. The theoretical U-quantile results of the previous setting apply to Q_n^α for any $0<\alpha<1$. The original Q_n proposed by Rousseeuw and Croux (1993) is defined as the $\binom{\lfloor n/2 \rfloor+1}{2}$ th order statistic of the $\binom{n}{2}$ values $|X_i - X_j|, 1 \le i < j \le n$. This corresponds roughly to $\alpha = 1/4$ and is aimed at achieving the maximal breakdown point of about 50%. However, a high-breakdown-estimator is counterproductive for change-point detection purposes: it is designed to discard a large portion of outliers, no matter how they are distributed spatially or temporarily. The perception of robustness in the change-point setting is conceptually different: we want to safeguard against a few outliers or several but evenly distributed over the observed sequence, as they may be generated by a heavy-tailed stationary process. A subsequence of outliers on the other hand, which exhibits some common characteristics, constitutes a structural change, which shall be detected rather than ignored.

The point may be illustrated by a small Monte Carlo simulation example: starting from an i.i.d. sequence of standard normal observations, we multiply the second half by some value λ. Table 1 compares rejection frequencies at the asymptotic 5% level of the test based on the original Q_n to those based on the sample variance and Gini's mean difference for sample sizes n = 60 and n = 500 and several values of λ (long-run variance estimation as in Section 5). In principle, if the sample size is large enough, the Q_n is able to pick up scale changes, i.e., the 1/4th quantile of all pairwise differences is larger than the 1/4th quantile of all pairs of the first half of the observations. For n = 500, this difference is sufficiently pronounced so that the test works. However, for n = 60, this difference is relatively small compared to the increased long-run variance if λ is large. This largely accounts for the decrease of power as λ increases. Additionally, the Q_n-based test grossly exceeds the size for small n. This may be attributed to a general bad "small sample behavior" of the test statistic. We observe a similar effect for the MAD-based test, cf. Section 5. Thus the Q_n is unsuitable for a quick detection of strong changes.

Accepted Manuscript

The problems can be overcome by considering Q_n^α for larger values of α than 1/4, and using such a scale estimator indeed leads to a workable change-point test also for n = 60. As a guideline for a suitable choice of α, we may look at the asymptotic efficiencies. Figure 1 plots the asymptotic relative efficiency of the Q_n^α at several scale families with respect to the respective maximum-likelihood estimator for scale. The solid line (normal distribution) is also depicted in Rousseeuw and Croux (1992). We are particularly interested in efficiency at heavy-tailed distributions and further include several members of the t_ν-family (for $\nu = 1/2, 1, 3, 10$) and the Laplace distribution. The latter has density $f(x) = 1/2 \exp(-|x|), x \in \mathbb{R}$. The mathematical derivations for this plot are given in Appendix B of the online supplement. The t_ν-distributions with $\nu = 1/2$ and ν = 1 are extremely heavy-tailed. We consider t_ν distributions with ν = 3 and ν = 5 to be more realistic data models, and these are also included in the simulation results presented in Section 5.

Altogether, Figure 1 suggests that $\alpha \approx 3/4$ may be a suitable choice as far as asymptotic efficiency is concerned. We ran simulations with many different values of α and found that Q_n^α generally performed best within the range $0.7 < \alpha < 0.9$. In the tables in Section 5, we report results for $\alpha = 0.8$.

3 Asymptotic results: null hypothesis and alternatives

In Section 3.1, we study the behavior of the studentized test statistics $\hat{D}_g^{-1}\hat{T}_g$ and $\hat{D}_Q^{-1}(\alpha)\hat{T}_Q(\alpha)$ under stationarity, thus deriving critical values for the respective tests. In Section 3.2, we investigate the asymptotic behavior of both test statistics under fixed alternatives, showing consistency of the tests against one-change alternatives.

3.1 Null hypothesis

We assume the data $X_1, \ldots, X_n$ to be a section of stationary process $(X_i)_{i \in \mathbb{Z}}$ with marginal distribution F satisfying the following assumption.

Accepted Manuscript

Assumption 3.1. *Let* $(X_i)_{i\in Z}$ *be a stationary process that is P-NED on an absolutely regular sequence* $(Z_n)_{n\in Z}$. *There is a* $\delta > 0$ *such that*

(a) the P-NED approximating constants a_l *and the absolute regularity coefficients* β_k *satisfy*

$$a_l \Phi\left(l^{-6}\right) = O\left(l^{-6\frac{2+\delta}{\delta}}\right) \text{ as } l \to \infty \qquad \text{and} \qquad \sum_{k=1}^{\infty} k\beta_k^{\frac{\delta}{2+\delta}} < \infty, \quad (6)$$

where Φ *may be any function satisfying condtion (2) in Definition 2.2, and*

(b) there is a positive constant *M* such that $E\,|\,X_0\,|^{2+\delta} \le M$ and $E\left|f_l\left(Z_{-l},\ldots,Z_l\right)\right|^{2+\delta} \le M$ for all $l \in \mathbb{N}$.

Then we have the following result about the asymptotic distribution of the test statistic $\hat{D}_g^{-1}\hat{T}_g$ based on Gini's mean difference. The proof is given in Appendix C.

Theorem 3.1. *If Assumptions 2.1 and 3.1 hold, then* $\hat{D}_g^{-1}\hat{T}_g \xrightarrow{\mathrm{d}} \sup_{0\le\lambda\le 1}\left|B(\lambda)\right|$, *where* $\left(B(\lambda)\right)_{0\le\lambda\le 1}$ *is a standard Brownian bridge.*

For the Q_n^{α}-based test, we require no moment condition, and it suffices that the short-range dependence condition (6) is satisfied for "$\delta = \infty$" (Assumption 3.2). However, instead of the moment condition we require a smoothness condition on *F* (Assumption 3.3).

Assumption 3.2. *Let* $(X_i)_{i\in Z}$ *be a stationary process that is P-NED on an absolutely regular sequence* $(Z_n)_{n\in Z}$ *such that the P-NED approximating constants* a_l *and the absolute regularity coefficients* β_k *satisfy* $a_l\Phi(l^{-6}) = O(l^{-6})$ *as* $l \to \infty$ *and* $\sum_{k=1}^{\infty} k\beta_k < \infty$, *where* Φ *is defined in Definition 2.2.*

Accepted Manuscript

Assumption 3.3. *The distribution F has a Lebesgue density f such that (a) f is bounded, (b) the support of f, i.e.,* $\overline{\{x \mid f(x) > 0\}}$ *, is a connected set, and (c) the real line can be decomposed in finitely many intervals such that f is continuous and (non-strictly) monotonic on each of them.*

We are now ready to state the following result concerning the asymptotic distribution of the Q_n^{α} -based change-point test statistic. The proof is given in Appendix C.

Theorem 3.2. *Under Assumptions 2.1, 2.2, 3.2, and 3.3, we have for any fixed* $0 < \alpha < 1$ *that* $\hat{D}_Q^{-1}(\alpha)\hat{T}_Q(\alpha) \xrightarrow{\mathrm{d}} \sup_{0 \le \lambda \le 1} |B(\lambda)|$ *, where* $\left(B(\lambda)\right)_{0 \le \lambda \le 1}$ *is a standard Brownian bridge.*

3.2 Fixed alternative: test consistency

In order to state the asymptotic behavior of the test statistics and fixed alternatives, we consider the following triangular array.

Assumption 3.4. *Let* $(X_i^{(n)})_{n \in \mathbb{N}, i \in \mathbb{Z}}$ *be such that*

$$X_i^{(n)} = \begin{cases} Y_i + \mu & \text{for } i \le [n\tau], \\ \lambda^\star Y_i + \mu & \text{for } i > [n\tau], \end{cases}$$

where $(Y_i)_{i \in \mathbb{Z}}$ *is a median-centered, stationary sequence,* $\mu \in \mathbb{R}$, $\lambda^\star > 0$ *, and* $\tau \in (0,1)$ *.*

We have the following result about the Gini's-mean-difference-based test statistic.

Theorem 3.3. *Let Assumption 3.4 hold with* $(Y_i)_{i \in \mathbb{Z}}$ *satisfying Assumption 3.1 and* $\lambda^\star \neq 1$ *. If further Assumption 2.1 holds, then* $\hat{D}_g^{-1}\hat{T}_g \xrightarrow{\mathrm{p}} \infty$ *as* $n \to \infty$ *.*

Accepted Manuscript

This implies that the test is consistent, i.e., $P(\hat{D}_g^{-1}\hat{T}_g > c)$ converges to 1 for any constant $c \in \mathbb{R}$. For the Q_n^α-based test, we need one further regularity assumption. Let

$$G^\star(t) = \tau^2 P(|Y_i - \tilde{Y}_j| \le t) + 2\tau(1-\tau)P(|Y_i - \lambda^\star \tilde{Y}_j| \le t) + (1-\tau)^2 P(\lambda^\star |Y_i - \tilde{Y}_j| \le t),$$

and $Q_\alpha^\star = \inf\left\{t \,\middle|\, G^\star(t) \ge \alpha\right\}$, where $\tilde{Y}_i$ is an independent copy of Y_i. Let F denote the marginal distribution of the process $(Y_i)_{i\in\mathbb{Z}}$.

Assumption 3.5. *Let $G^\star$ be differentiable in a neighborhood of $Q_\alpha^\star$, and the derivative is bounded away from 0. Furthermore, $Q_\alpha^\star \neq Q^\alpha$.*

Note that $P(|\lambda^\star Y_i - \lambda^\star \tilde{Y}_j| \le Q^\alpha) < P(|Y_i - \tilde{Y}_j| \le Q^\alpha)$ for $\lambda^\star > 1$. If additionally $P(|Y_i - \lambda^\star \tilde{Y}_j| \le Q^\alpha) \le P(|Y_i - \tilde{Y}_j| \le Q^\alpha)$, then $G^\star(Q^\alpha) < \alpha$ and consequently $Q_\alpha^\star > Q^\alpha$. The condition $P(|Y_i - \lambda^\star \tilde{Y}_j| \le Q^\alpha) \le P(|Y_i - \tilde{Y}_j| \le Q^\alpha)$ holds, e.g., if the density f of the distribution F is symmetric around $\mathrm{md}(F)$ and non-increasing to both sides. The next result implies consistency of the Q_n^α-based change-point test.

Theorem 3.4. *Let Assumption 3.4 hold with $(Y_i)_{i\in\mathbb{Z}}$ satisfying Assumptions 3.2 and 3.3, and $\lambda^\star \neq 1$. If further Assumptions 2.1, 2.2, and 3.5 hold, we have for any fixed $0 < \alpha < 1$ that $\hat{D}_Q^{-1}(\alpha)\hat{T}_Q(\alpha) \xrightarrow{p} \infty$ as $n \to \infty$.*

The proofs of Theorems 3.3 and 3.4 are given in Appendix C.

4 Practical aspects

4.1 Data-adaptive bandwidth selection

The rate conditions of Assumptions 2.1 on the HAC bandwidth b_n to achieve consistency of the long-run variance estimate are rather mild. However, the question remains how to choose b_n optimally for a given sequence of observations of length n. The answer depends on the degree of serial

Accepted Manuscript

dependence present in the sequence. Loosely speaking, choosing b_n too small may result in a size distortion, choosing it too large may render the tests conservative and less powerful. A common approach to this problem is to assume a parametric time-series model, minimize the mean squared error in terms of the parameters of the model and then plug-in estimates for the parameters. For instance, to estimate the long-run variance of an AR(1) process with autocorrelation parameter ρ, Andrews (1991, Section 6) gives an optimal bandwidth of

$$b_n = 1.447 n^{1/3} \left(\frac{4\rho^2}{(1-\rho^2)^2} \right)^{1/3} \qquad (7)$$

if the Bartlett kernel is used. Alternatively, non-parametric bandwidth selection schemes, based on the inspection of the whole autocorrelation function, have been considered. The basic idea is to find a maximal lag after which the autocorrelations may be regarded as negligible. Politis (2003, 2011) proposes the following: Letting $\hat{\rho}_j$ denote the sample autocorrelation for lag j, a maximal lag $l_{\max}$ is selected as the smallest non-negative integer k such that $\max\{|\hat{\rho}_k|,\ldots,|\hat{\rho}_{k+\kappa_n}|\} \le 2\sqrt{\log_{10}(n)/n}$ for $\kappa_n = \max\{5, \sqrt{\log_{10}(n)}\}$. This is used in connection with a flat-top kernel W, where $c_{\text{eff}} = \sup\{x \,|\, W(x) \ge 0.99\}$ describes the range in which W_n is "effectively" 1. The empirical bandwidth $\hat{b}_n$ is then $l_{\max}$ multiplied by c_{eff}^{-1} so that the "effective flat-top range" of the kernel stretches until lag $l_{\max}$. In the simulations we use the kernel

$$W(x) = \begin{cases} 1 & |x| < 1/2, \\ \{1 - 4(|x| - 1/2)^2\}^2 & 1/2 \le |x| < 1, \\ 0 & 1 \le |x| \end{cases}$$

with $c_{\text{eff}} = 0.536$, which can be viewed as a smoothed version of the trapezoidal flat-top kernel. We adapt this bandwidth-selection rule to the current setting in two ways:

Accepted Manuscript

(1) The same procedure is applied to the autocorrelation of the squared centered data, resulting in an analogous maximal lag $l_{\max,2}$, and the maximum of $l_{\max}$ and $l_{\max,2}$ is taken. If the serial dependence is such that the observations are uncorrelated, but the squared observations correlated (as in GARCH models), this will generally affect the long-run variance of scale estimators.

(2) The maximal lag is limited to $n^{1/3}$. Such an upper limit is a crucial adjustment for change-point tests. Maybe more important than an accurate estimation of the long-run variance under stationarity is the prevention of a too-strong inflation in the presence of a change. A change-point may lead to very persistent autocorrelations (a change in scale primarily affects the autocorrelations of the squares), hence to large values of $l_{\max}$ according to the above rule, and hence to very large long-run variance estimates.

4.2 Bootstrap

An alternative way of assessing critical values for the tests is by means of bootstrap methods. A variety of bootstrap procedures have been proposed for dependent data, e.g., the block bootstrap (Künsch, 1989), the stationary bootstrap (Politis and Romano, 1994), the tapered block bootstrap (Paparoditis and Politis, 2001), or the dependent wild (or multiplier) bootstrap (Shao, 2010). The block bootstrap for U-statistics (such as Gini's mean difference) has been studied by Dehling and Wendler (2010). Recently, Leucht and Neumann (2013) and Bücher and Kojadinovic (2016b) showed the consistency of the dependent multiplier bootstrap for U-statistics and also established the validity of dependent multiplier bootstrap procedures for change-point test statistics based on this class of statistics. For quantiles, the block bootstrap was investigated by Sun and Lahiri (2006) and by Sharipov and Wendler (2013) and the multiplier bootstrap by Doukhan et al. (2015). For U-quantiles, such as the Q_n^{α}, we are unaware of any work concerning bootstrap methods. We conjecture that the overlapping block bootstrap is consistent for this class of statistics as well, and the simulation results in

Accepted Manuscript

Section 5 provide evidence of its validity, but a proof goes beyond the scope of the current paper.

We use the dependent block bootstrap in our simulations, since it is easy to implement. Furthermore, the dependent wild (or multiplier) bootstrap might lead to a possible non-monotonous bootstrap version of the distribution function of the pairwise difference, so it is not straightforward to define a bootstrap version of Q_n^{α}, the α-quantile of the bootstrapped pairwise difference.

The (overlapping, non-circular) dependent block bootstrap with blocklength l proceeds as follows: A bootstrap sample $(X_1^{\star},\ldots,X_{kl}^{\star})$ of length $k \cdot l$, where $k = \lfloor n/l \rfloor$, is created by selecting $J_1,\ldots,J_k$, i.i.d., uniformly from $\{1,\ldots,n-l+1\}$, and concatenating the blocks $(X_{J_i},\ldots,X_{J_i+l-1})$ for $1 \le i \le k$. Thus, for any $0 \le i \le k-1$, we have

$$P^{\star}\left[(X_{il+1}^{\star},\ldots,X_{(i+1)l}^{\star}) = (X_j,\ldots,X_{j+l-1})\right] = \frac{1}{n-l+1}$$

for all $j = 1,\ldots,n-l+1$, where $P^{\star}$ denotes the bootstrap distribution conditionally on $(X_i)_{1\le i\le n}$. Then, any of test statistics $\hat{T}_s$ is applied to the bootstrap sample, the procedure is repeated N times, and the level-α critical value for the test is obtained as the empirical $(1-\alpha)$-quantile of the thus obtained N bootstrap test statistics.

When carrying out the dependent block bootstrap (similarly for any other dependent bootstrap scheme), one faces with the blocklength selection a very similar challenge to the bandwidth selection problem in case of the long-run variance estimation. This problem has also been widely studied. For instance, Carlstein (1986) considered a block subsampling scheme and obtained

$$l_n = 1.587\left(\frac{\rho^2}{(1-\rho^2)^2}\right)^{1/3} n^{1/3} \quad (8)$$

Accepted Manuscript

as an optimal blocklength in the AR(1) setting with autocorrelation -parameter ρ. As overlapping sub-sampling gives the same variance estimate as the overlapping block bootstrap, we may adopt this idea to propose a data-adaptive blocklength selection rule: estimate ϕ by the lag-1 autocorrelation of $(X_i - \bar{X})^2, i = 1, \ldots, n$ for $\hat{T}_{\sigma^2}$, by the lag-1 autocorrelation of $|X_i - \mathrm{md}(F_n)|$ for $\hat{T}_d$, by the autocorrelation of $\hat{\phi}_n(X_i)$ for $\hat{T}_g$, and by the autocorrelation of $\hat{\xi}_n(X_i)$ for $\hat{T}_Q$. (Lower and upper bounds for the blocklength are advisable here as well.) This seems justified, as the optimal block length for the for *U*-statistics is known to be asymptotically equivalent to the optimal block length for the linear part, see Dehling and Wendler (2010) and Bücher and Kojadinovic (2016a). The optimal block length for more general bootstrap methods and more general processes where studied by Politis and White (2004) (see also Patton et al. (2009)). This technique is adapted, e.g., by Kojadinovic et al. (2015) for a change-point test based on Spearman's rho.

Alternatively, a nonparametric blocklength selection rule based on the whole autocorrelation function, similar to the rule discussed in Section 4.1, may be applied here. In fact, empirical bandwidths selection and empirical blocklength selection are intrinsically linked. In both cases, a maximal lag of non-negligible serial dependence is to be determined. For comparability reasons, we use the same maximal lag $l_{\max}$ from Section 4.1 also in the simulations for the block bootstrap.

5 Simulations

We investigate the empirical size and power of the change-point tests based on the test statistics introduced in Section 2.2. Simulation results with long-run variance estimation (Section 5.1) and with dependent block bootstrap (Section 5.2) are presented.

5.1 Simulations with long-run variance estimation

The data $X_1, \ldots, X_n$ are generated as

Accepted Manuscript

$$X_i = \begin{cases} Y_i, & 1 \le i \le \lfloor \tau n \rfloor, \\ \lambda Y_i, & \lfloor \tau n \rfloor < i \le n, \end{cases}$$

for some $0 < \tau < 1$, where $(Y_i)_{i \in Z}$ is a stationary process. Thus we have three general parameters: the size of the change λ, the (relative) location of the change τ, and the sample size n. The tables below report results for $\lambda = 1$ (null hypothesis) and $\lambda = 2$ (alternative), $\tau = 1/4$, 1/2, and 3/4, and $n = 60$, 120, 240, and 500. A fourth "parameter" is the process $(Y_i)_{i \in Z}$, particularly the type of serial dependence and the tailedness of its marginal distribution. For the former, we consider two scenarios:

Scenario A. Linear process: $(Y_i)_{i \in Z}$ follows an AR(1) process $Y_i = \rho Y_{i-1} + \epsilon_i, i \in Z$, for some $-1 < \rho < 1$, where the ϵ_i are an i.i.d. sequence with a mean-centered distribution. We present results for $\rho = 0$ and $\rho = 0.8$ and for five different innovation distributions: the standard normal $N(0, 1)$, the standard Laplace $L(0, 1)$, the normal scale mixture $NM(\gamma, \varepsilon)$ with $\gamma = 3$, $\varepsilon = 0.01$, and t_ν-distributions with $\nu = 3$ and $\nu = 5$. The latter four are, to varying degrees, heavier-tailed than the normal. The normal mixture distribution $NM(\gamma, \varepsilon)$ has density $f(x) = (\varepsilon/\gamma)\phi(x/\gamma) + (1-\varepsilon)\phi(x), x \in R$, where ϕ denotes the standard normal density. It captures the notion that the majority of the data stems from the normal distribution, except for some small fraction, which stems from a normal distribution with a γ times larger standard deviation. This model has been popularized by Tukey (1960), who argued that $\gamma = 3$ is a realistic value in practice, and pointed out that in this case the mean deviation d_n is more efficient than the standard deviation for ε-values as low as 1%.

Scenario B. GARCH model: $(Y_i)_{i \in N}$ follows the GARCH(1,1) process

$$\begin{cases} Y_i = \sigma_i \epsilon_i, \\ \sigma_i^2 = a_0 + a_1 Y_{i-1}^2 + b\sigma_{i-1}^2, \qquad i \in Z, \end{cases}$$

with $a_0 = 0.1, a_1 = 0.85$, and $b_1 = 0.05$, where the ϵ_i, $i \in Z$, are i.i.d. standard normal. These parameter settings are observable in financial data (e.g., log

Accepted Manuscript

returns of price indices) and tend to generate quite pronounced, but relatively short volatility clusters. Such a model may be regarded as an outlier-generating process, where the outliers have a stronger tendency to cluster as compared to a heavy-tailed i.i.d. sequence. Furthermore, the Y_i themselves are uncorrelated, but their squares are correlated.

Regarding the long-run variance estimation, we take the flat-top kernel W described in Section 4.1 for the HAC estimation, and the Epanechnikov kernel $K(t) = (3/4)(1-t^2)\ 1_{[-1,1]}(t)$ for the density estimation. Further, the bandwidths h_n (for the density estimation) and b_n (HAC estimation) are chosen data-dependent: $h_n = I_n n^{-1/3}$, where I_n denotes the sample interquartile range of the data points the kernel density estimator is applied to, and b_n according to the empirical rule described in Section 4.1. We found the results of the long-run variance estimation to differ little with respect to the choice of kernels. The bandwidth b_n is chosen the same for all estimators. Since b_n plays a very similar role in all long-run variance estimators, this allows a fair comparison and puts the focus on the impact of the different estimators. For each setting we generate 1000 repetitions. All tables report empirical rejection frequencies (in %) at the asymptotic 5% significance level, i.e., we count how often the test statistics exceed 1.358, i.e., the 95%-quantile of the limiting distribution of the studentized test statistics under the null.

Analysis of size. Table 2 reports results for Scenario A for six different change-point tests based on the variance σ_n^2 (Var), the mean deviation d_n (MD), Gini's mean difference g_n (GMD), the median absolute deviation m_n (MAD), the original Q_n, and the $Q_n^{0.8}$. The Q_n and the MAD heavily exceed the size. This effect wears off as n increases, but rather slowly. The Q_n shows an acceptable size behavior for n = 500, the MAD not even for this sample size. This behavior can be described as a discretization effect. A similar observation is made at the median-based change-point test for location, which is discussed in detail in Vogel and Wendler (2017, Section 4). Due to the size distortion, the MAD and the Q_n are excluded from any further power analysis. Also the other tests show some size exceedance in the AR(1) case for n = 60,

Accepted Manuscript

most notably the $Q_n^{0.8}$. The size results for Scenario B are reported in Table 5. Most tests appear to be conservative.

Analysis of power. Tables 3 and 4 list empirical rejection frequencies under the alternatives in Scenario A: Table 3 for serial independence ($\rho = 0$) and Table 4 for strong serial dependence ($\rho = 0.8$). In both cases, the scale changes by factor $\lambda = 2$. We make the following observations:

(1) All tests have better power at independent sequences than dependent sequences. Note that $\rho = 0.8$ constitutes a scenario of rather strong serial dependence.

(2) All tests loose power as the tails of the innovation distribution increase, but the loss is much more pronounced for the variance than for the other estimators. The distributions listed in the tables are in ascending order according to their kurtosis. The kurtoses of $N(0, 1)$, $NM(3,1\%)$, $L(0, 1)$, t_5, and t_3 are 0, 1.63, 3, 6, and ∞, respectively.

(3) The tests generally have a higher power for $\tau = 3/4$ than for $\tau = 1/4$. This may appear odd since in both cases the change occurs equally far away from the center of the sequence. However, since we consider changes from a smaller ($\lambda = 1$) to a larger scale ($\lambda = 2$), a sequence with $\tau = 1/4$ has a higher overall variability and hence yields a larger long-run variance estimate than a sequence with $\tau = 3/4$. Since we divide the test statistics by the (root of the) long-run variance estimates, this implies a difference in power.

(4) The GMD-based test turns out to have the overall best performance, with $Q_n^{0.8}$ and MD not trailing far behind. Based on the simulation results for size and power, the $Q_n^{0.8}$ can be seen to provide a sensible change-point test for scale, but some caution should be taken for sample sizes below $n = 100$.

Accepted Manuscript

(5) It is interesting to note that the GMD, the $Q_n^{0.8}$, and the MD dominate the variance-based test also under normality. One explanation is that a change in scale tends to blow up the long-run variance estimate $\hat{D}_{\sigma^2}$ much more than the corresponding estimates for the other estimators. To illustrate this, consider a sequence of i.i.d. *N*(0, 1)-variates $X_1,\ldots,X_{\lfloor n/2\rfloor}$ followed by a sequence of i.i.d. *N*(0, 4)-variates $X_{\lfloor n/2\rfloor+1},\ldots,X_n$. For large *n*, the quantity that $\hat{D}^2_{\sigma^2}$ estimates when applied to $X_1,\ldots,X_n$ can be seen to be $ASV(\sigma_n^2; NM(\gamma=2,\epsilon=1/2)) = E(Y^4)-E(Y^2)^2$ for $Y \sim NM(\gamma=2,\epsilon=1/2)$, whereas $\hat{D}^2_d$ estimates $ASV(d_n; NM(2,1/2)) = E(Y^2)-(E|Y|)^2$ with *Y* as before. Compared to a stationary sequence of *N*(0, 1)-variates, the former quantity is blown up by the factor 19.25, the latter only by 2.93.

Table 5 contains power results for the GARCH(1,1) model of Scenario B. GMD, MD, and $Q_n^{0.8}$ show similar powers and outperform the variance test. All tests have a lower power than in the independence as well as the AR(1) setting. However, the lower size at the GARCH model must be taken into account. Conditional heteroscedasticity is a challenging setting for any change-in-scale test, and the tests generally cope well.

5.2 Simulations with bootstrapping

For the bootstrap simulations, we examine the same Scenarios A (independence and AR) and B (GARCH) as before, but restrict our attention to sample sizes *n* = 60, 120, 240 and three marginal distributions: *N*(0, 1), *L*(0, 1), and t_3. We apply the overlapping dependent block bootstrap, as described in Section 4.2, with data-adaptive blocklength selection analogous to the automatic bandwidth selection described in Section 4.1. We take $l_{\max}$ (from Section 4.1) as the blocklength for bootstrapping all test statistics. We use 1000 bootstrap samples and, as before, 1000 repetitions for each setting. Size results under the null hypothesis for Scenario A are presented in Table 6, power results in Table 7. Results for size and power for Scenario B are given in Table 8. The numbers are qualitatively comparable to the simulations with

Accepted Manuscript

long-run variance estimation. We arrive at the same conclusions regarding the ranking of the estimators. In particular, we observe the following:

Analysis of size. We have noticed that the studentized tests exceed the size for $n = 60$ in the AR setting (Table 2, right columns), which prompts the question if this is improved by bootstrapping. The answer is generally no. With the bootstrap, the size exceedance is of smaller magnitude for $n = 60$, but, contrary to studentized tests, it appears to be persistent also for larger sample sizes. The bootstrapped tests also slightly exceed the size in the GARCH setting. For i.i.d., sequences, bootstrapping and studentization keep the size also for $n = 60$. The $Q_n^{0.8}$ shows the best overall size behavior with the bootstrap.

Analysis of power. The bootstrapped tests generally have a higher power than their studentized counterpart. The power gain is marginal under independence, but quite substantial for the dependent sequences (AR and GARCH). However, this must be put in perspective with the size exceedance of the bootstrap. Adjusting the critical values of the studentized tests such that they exhibit the same rejection frequency under the null hypothesis would result in similar powers under alternatives.

Ultimately it is difficult to arrive at a definite decision whether to prefer studentization or bootstrapping. A conservative recommendation would be to bootstrap for sample sizes below $n = 100$ and to studentize for larger sample sizes.

6 Data Example

We consider two data examples: a hydrological and a financial time series. The first data set consists of the annual maximum discharge (in cubic meters per second) of the river Rhine at Cologne, Germany, in the years 1817 to 2009 ($n = 193$). The time series is plotted in Figure 2, top row. Figure 3 depicts a normal q-q plot, which reveals that the marginal distribution is fairly normal. Furthermore, the autocorrelation function of the data and the squared mean-centered data are plotted. They indicate a weak serial dependence. We

Accepted Manuscript

thus apply the change-point tests from the previous section with long-run variance estimation settings as before except for b_n = 4. The latter is in consistency with the other data example. The results are very similar for smaller bandwidths.

The change-point test processes $\hat{D}_s^{-1}\left(k/\sqrt{n}\,|\,s_{1:k}-s_{1:n}\,|\right)_{2\le k\le n}$ show a fair agreement for $s_n=\sigma_n^2, d_n, g_n$, and $Q_n^{0.8}$, cf. Figure 2, middle and bottom row. All attain their maxima at 1919 with p-values ranging from 0.021 to 0.046, i.e., they confirm the existence of a change in scale around 1920, which is suggested by a visual inspection of the series. This change coincides with the implementation of a variety of structural river works upstream from Cologne, particularly along the Rhine's tributaries Main and Neckar in the early 1920s. For illustration purposes, we also plot the corresponding curves for the original Q_n ($\alpha=1/4$) and the MAD in the bottom row of Figure 2. Both are very rugged, distinctively different from the other curves, and yield p-values above 5%.

The second example consists of the log returns of the daily closings of the Volkswagen share, traded at the German stock exchange in Frankfurt, within the last three quarters of the year 2001 (n = 196). The impact of 9/11 on the volatility of the series is clearly visible, cf. Figure 4, top row. This example differs in a variety of features from the first example, which shall illustrate the differences of the tests. The normal q-q plot shows heavier than normal tails, and the autocorrelation function of the squared sequence reveals some serial dependence (cf. Figure 5). Furthermore, there is a short period of strong oscillation from July 10–12, 2001 (the reason for which is not known to us). Removing these three dates from the series, all tests consistently detect a change with p-values of 1% and less and place it at the beginning of September. However, *with* those three days, the variance-based change-point test process $\hat{D}_{\sigma^2}^{-1}\left(k/\sqrt{n}\,|\,\sigma_{1:k}^2-\sigma_{1:n}^2\,|\right)_{2\le k\le n}$ attains its maximum at July 5 and yields a p-value of 0.31. This is an example where few outliers mask an apparent change. The tests based on MD and GMD behave in principle

Accepted Manuscript

similarly, but provide some mild evidence for a change with p-values of about 0.10. The $Q_n^{0.8}$, however, gives a p-value of 0.03, and the maximum is attained at September 7.

The curves for the MAD and Q_n are plotted as well. Both curves, again, look distinctively different from the others, and the respective tests do not provide strong evidence for a change. The tests were carried out, as before, with b_n = 4 and the other long-run variance estimation parameters as in Section 5. All simulations and data analyses were performed in R (R Core Team, 2015).

7 Conclusion and Outlook

We have studied the problem of detecting changes in scale beyond the established sum-of-squares methodology. We have considered test statistics based on alternative scale measures, which have a better outlier-resistance and a better efficiency at heavy-tailed distributions than the sample variance. The MAD and the original Q_n (i.e., Q_n^{α} with about $\alpha = 1/4$) may be confidently discarded for the purpose of change-point detecting, whereas the mean deviation, Gini's mean difference, and the Q_n^{α} for $0.7 < \alpha < 0.9$ provide very good alternatives. They improve upon the classical test not only under heavy tails but also under normality. We found Gini's mean difference and the Q_n^{α}, which are both based on pairwise differences, to altogether outperform the mean deviation. Our general recommendation is to use Gini's mean difference in case of normal or near-normal data situations, and to use the $Q_n^{0.8}$ or $Q_n^{0.75}$ if the occurrence of gross errors is suspected.

An alternative way of robustifying the classical methodology was studied by Lee and Park (2001), who considered a version of the σ^2-based test with truncated observations. Their simulation results suggest quite a substantial loss in power under normality, whereas we observe the opposite for our robustification approach. There is another conceptual advantage of pairwise differences: there is no location to estimate. At skewed distributions, taking, e.g., the mean or the median leads to distinctively different scale estimators. This ambiguity does not exist for pairwise-difference-based estimators.

Accepted Manuscript

Gini's mean difference and the Q_n^{α} are based on the kernel $g(x,y)=|x-y|$ of order two. We have noted that both have a considerably higher efficiency under normality as compared to the respective estimators based on distances to the central location. So, may kernels of higher order even be better? A general observation seems to be that this is not the case, see, e.g., Rousseeuw and Croux (1992, Section 4). The higher computational cost of higher-order-kernel methods tends to be not justified by a better efficiency or robustness.

A crucial part of all change-point tests for dependent data is the estimation of the long-run variance. We have proposed estimators based on HAC kernel estimation, which is common in the change-point context. Another estimation technique is block sub-sampling, see, e.g., Dehling et al. (2013) and the references therein. Instead of studentizing the test statistic, i.e., dividing by a (long-run) variance estimate, bootstrapping the test statistic has become a very popular alternative for obtaining critical values. We have implemented an overlapping block bootstrap, but other bootstrapping schemes are likewise possible. An entirely different approach, which avoids any unknown scaling constants in the limit distribution of the test statistic is the self-normalization approach as proposed by Shao and Zhang (2010).

Finally, high-breakdown-point estimators such as the MAD and the original Q_n have also been studied and turned out to be rather unsuited for the problem at hand. This leads to the question of what type of 'optimal' robustness can be achieved in a change-point context, and if this can be mathematically formalized.

Accepted Manuscript

Acknowledgement

The authors were supported by the Collaborative Research Centre 823 *Statistical modelling of nonlinear dynamic processes* and the Konrad-Adenauer-Stiftung. The authors thank Svenja Fischer for the river Rhine discharge data set, Marco Thiel for the stock exchange data set, and Silke Henkes, whose knowledge of Linux helped the simulations to finish two times

faster. Their gratitude is extended to the referees and editors, whose valuable comments greatly improved the paper.

References

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *Annals of Statistics*, 37(6B):4046–4087.

Bücher, A. and Kojadinovic, I. (2016a). A dependent multiplier bootstrap for the sequential empirical copula process under strong mixing. *Bernoulli*, 22(2):927–968.

Bücher, A. and Kojadinovic, I. (2016b). Dependent multiplier bootstraps for non-degenerate U-statistics under mixing conditions with applications. *Journal of Statistical Planning and Inference*, 170:83–105.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics*, 14(3):1171–1179.

de Jong, R. M. and Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, 68(2):407–424.

Dehling, H., Fried, R., Sharipov, O. S., Vogel, D., and Wornowizki, M. (2013). Estimation of the variance of partial sums of dependent processes. *Statistics & Probability Letters*, 83(1):141–147.

Dehling, H., Vogel, D., Wendler, M., and Wied, D. (2017). Testing for changes in Kendall's tau. *Econometric Theory*, 33(6):1352–1386.

Accepted Manuscript

Dehling, H. and Wendler, M. (2010). Central limit theorem and the bootstrap for U-statistics of strongly mixing data. *Journal of Multivariate Analysis*, 101(1):126–137.

Doukhan, P., Lang, G., Leucht, A., and Neumann, M. H. (2015). Dependent wild bootstrap for the empirical process. *Journal of Time Series Analysis*, 36(3):290–314.

Gerstenberger, C. and Vogel, D. (2015). On the efficiency of Gini's mean difference. *Statistical Methods & Applications*, 24(4):569–596.

Gombay, E., Horváth, L., and Husková, M. (1996). Estimators and tests for change in variances. *Statistics & Risk Modeling*, 14(2):145–160.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. Wiley, 2nd edition.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

Inclan, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923.

Kojadinovic, I., Quessy, J.-F., and Rohmer, T. (2015). Testing the constancy of Spearman's rho in multivariate time series. *Annals of the Institute of Statistical Mathematics*, 65(5):292–954.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241.

Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80(391):736–741.

Accepted Manuscript

Lee, S. and Park, S. (2001). The cusum of squares test for scale changes in infinite order moving average processes. *Scandinavian Journal of Statistics*, 28(4):625–644.

Leucht, A. and Neumann, M. H. (2013). Dependent wild bootstrap for degenerate U-and V-statistics. *Journal of Multivariate Analysis*, 117:257–280.

Maronna, R. A., Martin, D. R., and Yohai, V. J. (2006). *Robust statistics: Theory and methods.* Wiley Series in Probability and Statistics. Chichester: Wiley.

Nair, U. S. (1936). The standard error of Gini's mean difference. *Biometrika*, 28:428–436.

Paparoditis, E. and Politis, D. N. (2001). Tapered block bootstrap. *Biometrika*, 88(4):1105–1119.

Patton, A., Politis, D. N., and White, H. (2009). Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. *Econometric Reviews*, 28(4):372–375.

Politis, D. N. (2003). Adaptive bandwidth choice. *Journal of Nonparametric Statistics*, 15(4-5):517–533.

Politis, D. N. (2011). Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. *Econometric Theory*, 27(4):703–744.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Politis, D. N. and White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, 23(1):53–70.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Accepted Manuscript

Rousseeuw, P. J. and Croux, C. (1992). Explicit scale estimators with high breakdown point. In Dodge, Y., editor, *L1-Statistical analysis and related methods*, volume 1, pages 77–92. North-Holland.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235.

Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association*, 105(491):1228–1240.

Sharipov, O. S. and Wendler, M. (2013). Normal limits, nonnormal limits, and the bootstrap for quantiles of dependent data. *Statistics & Probability Letters*, 83(4):1028–1035.

Sun, S. and Lahiri, S. N. (2006). Bootstrapping the sample quantile of a weakly dependent sequence. *Sankhyā: The Indian Journal of Statistics*, pages 130–166.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin et al., editor, *Contributions to Probability and Statistcs. Essays in Honor of Harold Hotteling*, pages 448–485. Stanford University Press.

Vogel, D. and Wendler, M. (2017). Studentized U-quantile processes under dependence with applications to change-point analysis. *Bernoulli*, 23(4B):3114–3144.

Wied, D., Arnold, M., Bissantz, N., and Ziggel, D. (2012). A new fluctuation test for constant variances with applications to finance. *Metrika*, 75(8):1111–1127.

Accepted Manuscript

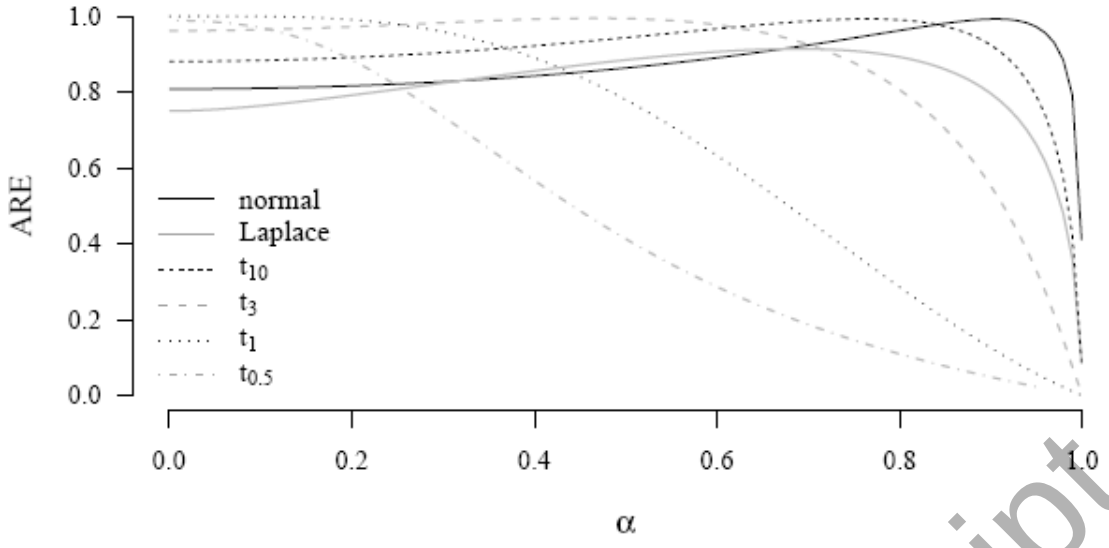

Fig. 1 Asymptotic relative efficiencies (AREs) of Q_n^{α} at normal, Laplace, and several *t* distributions wrt respective maximum-likelihood estimators of scale.

Accepted Manuscript

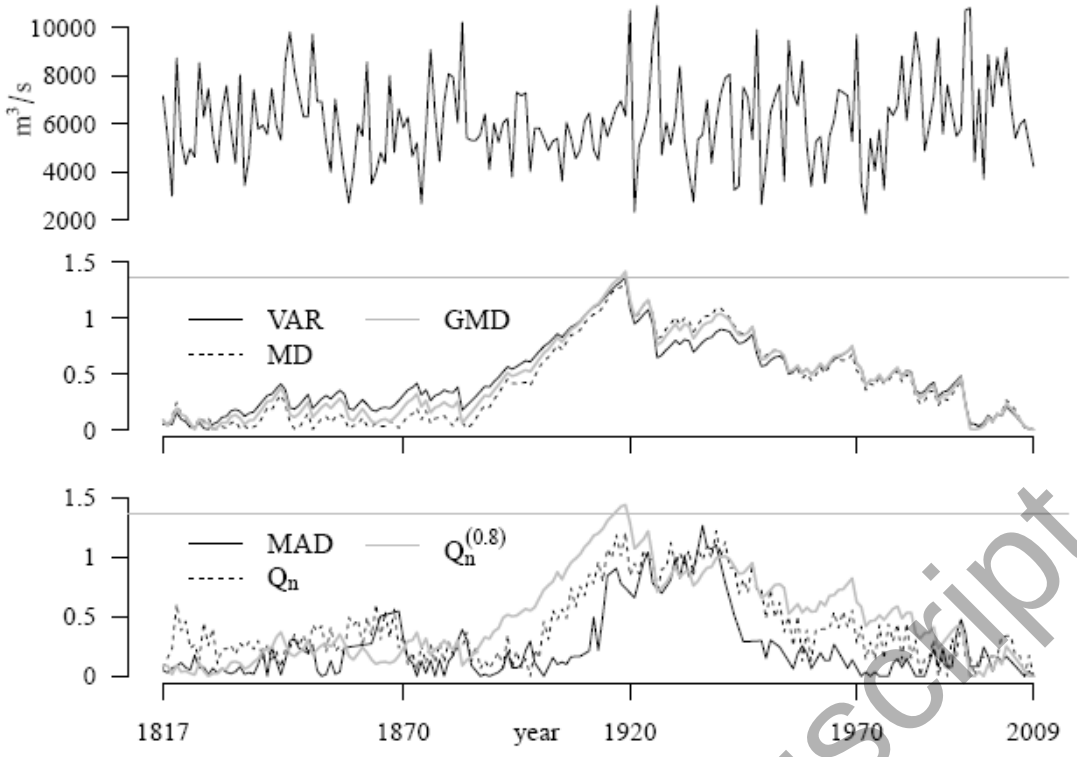

Fig. 2 Top row: annual maximum discharge (m^3 / s) of the river Rhine at Cologne, Germany, between 1817 and 2009. Middle and bottom rows: change-point processes $\hat{D}_s^{-1}\left(k/\sqrt{n} \mid s_{1:k} - s_{1:n} \mid\right)_{2\le k\le n}$ for estimators $s_n = \sigma_n^2$, d_n, g_n (middle) and m_n, Q_n, and $Q_n^{0.8}$ (bottom); HAC kernel bandwidth: $b_n = 4$.

Accepted Manuscript

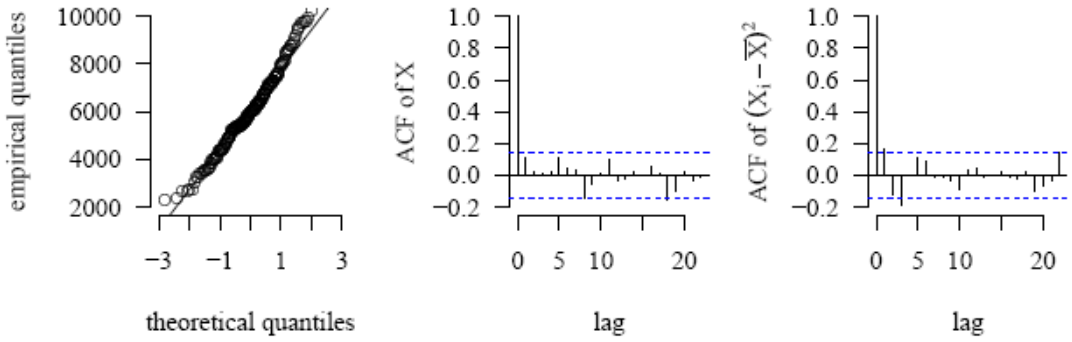

Fig. 3 Hydrology data from Figure 2: Normal q-q plot, ACF of data, and ACF of squared centered data.

Accepted Manuscript

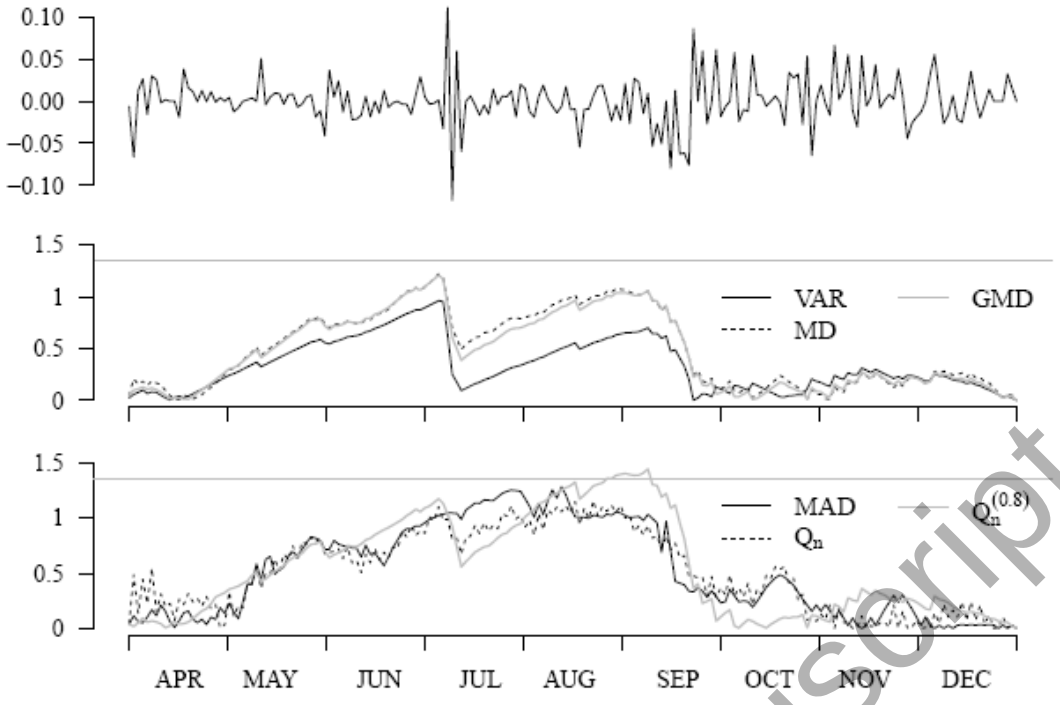

Fig. 4 Top row: daily log returns of VW share from April to December 2001. Middle and bottom rows: change-point processes $\hat{D}_s^{-1}\left(k/\sqrt{n}\,|\,s_{1:k}-s_{1:n}\,|\right)_{2\le k\le n}$ for estimators $s_n=\sigma_n^2$, d_n, g_n (middle) and m_n, Q_n, and $Q_n^{0.8}$ (bottom); HAC kernel bandwidth: $b_n = 4$.

Accepted Manuscript

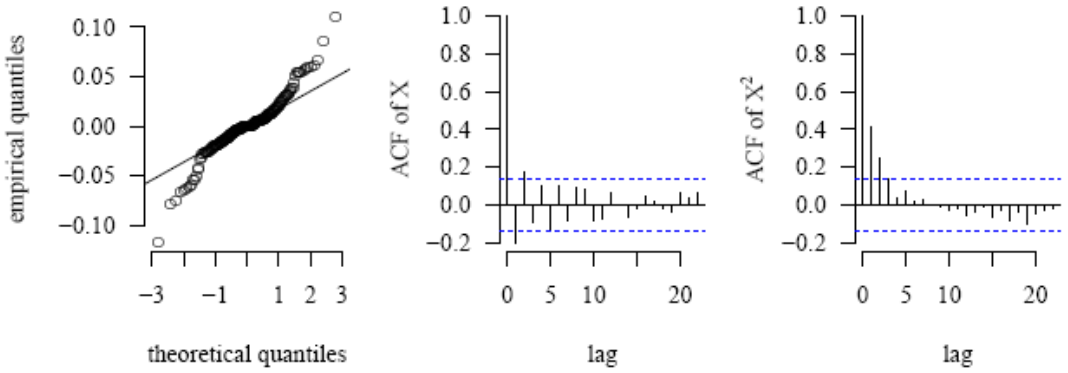

Fig. 5 Financial data from Figure 4: Normal q-q plot, ACF of data, and ACF of squared data.

Accepted Manuscript

Table 1 Power of change-point tests at asymptotic 5% level, based on sample variance/Gini's mean difference/Q_n for independent, centered normal observations. Standard deviation changes from 1 in first half to λ in second half. (Note that this refers to $Q_n^{0.25}$; for much more positive results on $Q_n^{0.8}$, see Tables 2–8.)

	standard deviation λ in second half				
sample size	1.0	1.2	1.5	2.0	3.0
n = 60	.04/.04/.44	.04/.07/.32	.12/.24/.18	.23/.51/.06	.36/.82/.01
n = 500	.04/.04/.08	.63/.64/.52	1/1/1	1/1/1	1/1/1

Accepted Manuscript

Table 2 *Test size for Scenario A.* Rejection frequencies (%) of change-point tests with long-run variance estimation based on six different scale estimators at stationary sequences (independent sequence and AR(1) with $\rho = 0.8$). Asymptotic 5% significance level; sample sizes $n = 60, 120, 240, 500$; five different innovation distributions.

	independence						*AR*(1) with $\rho = 0.8$					
Estimator	Var	MD	GMD	MAD	Q_n	$Q_n^{0.8}$	Var	MD	GMD	MAD	Q_n	$Q_n^{0.8}$
n = 60												
$N(0,1)$	4	4	4	20	38	5	8	8	12	25	17	13
$NM(3,0.01)$	4	5	4	18	39	7	9	7	13	24	14	14
$L(0,1)$	4	3	4	19	32	5	8	7	10	23	12	13
t_5	4	5	4	21	37	7	9	8	13	23	14	13
t_3	3	6	3	19	32	8	6	6	10	24	10	10
n = 120												
$N(0,1)$	4	5	4	16	22	5	4	5	7	22	7	6
$NM(3,0.01)$	3	4	3	16	21	4	5	3	7	18	6	6
$L(0, 1)$	3	4	3	14	20	4	4	5	8	18	6	7
t_5	3	4	2	14	19	6	4	5	7	19	7	7
t_3	2	4	2	18	21	8	3	3	5	18	7	4
n = 240												
$N(0,1)$	3	3	3	13	10	4	2	3	5	14	4	4
$NM(3,0.01)$	3	4	4	11	12	4	4	5	5	16	4	5
$L(0,1)$	2	4	3	9	9	4	2	3	3	14	4	4
t_5	3	4	3	13	12	4	2	2	4	14	3	4
t_3	2	2	2	12	11	5	2	2	3	13	4	4
n = 500												
$N(0,1)$	5	5	4	10	7	4	3	3	3	10	3	3
$NM(3,0.01)$	4	6	4	11	9	5	2	2	3	11	3	3

Accepted Manuscript

	independence						$AR(1)$ with $\rho = 0.8$					
$L(0, 1)$	4	5	4	10	6	4	3	3	4	10	4	3
t_5	3	4	3	10	8	4	2	3	3	11	4	3
t_3	2	4	4	10	7	7	2	4	4	10	4	5

Accepted Manuscript

Table 3 *Test power for Scenario A.* Rejection frequencies (%) at asymptotic 5% level. Change-point tests with long-run variance estimation based on variance (Var), mean deviation (MD), Gini's mean difference (GMD), and $Q_n^{0.8}$. **Independent observations**; scale changes by **factor** $\lambda = 2.0$.

Change location:	[*n*/4]				[*n*/2]				[3*n*/4]			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
n = 60												
$\mathcal{N}(0,1)$	13	21	32	32	59	64	76	66	52	49	60	46
$NM(3,0.01)$	12	19	30	31	50	61	70	62	48	48	56	43
$\mathcal{L}(0,1)$	5	9	14	13	24	40	46	35	26	32	36	25
t_5	6	12	18	17	30	49	55	44	28	37	42	29
t_3	4	8	11	12	16	31	33	31	19	25	29	22
n = 120												
$\mathcal{N}(0,1)$	47	68	79	75	95	96	98	96	92	90	93	90
$NM(3,0.01)$	38	61	70	72	86	96	97	97	84	89	92	88
$\mathcal{L}(0,1)$	15	36	40	33	63	86	85	74	57	68	70	56
t_5	17	44	47	46	67	91	90	87	64	77	79	72
t_3	9	30	30	28	37	74	71	73	36	60	59	56
n = 240												
$\mathcal{N}(0,1)$	97	99	100	100	100	100	100	100	100	100	100	100
$NM(3,0.01)$	80	99	99	99	96	100	100	100	97	100	100	99
$\mathcal{L}(0,1)$	49	87	87	79	95	99	99	99	93	97	97	94
t_5	49	93	91	92	91	100	99	100	89	98	98	98
t_3	16	71	62	74	61	96	94	98	60	89	86	94
n = 500												
$\mathcal{N}(0,1)$	100	100	100	100	100	100	100	100	100	100	100	100
$NM(3,0.01)$	96	100	100	100	100	100	100	100	100	100	100	100

Accepted Manuscript

Change location:	[n/4]				[n/2]				[3n/4]			
L(0,1)	94	100	100	100	100	100	100	100	100	100	100	100
t_5	87	100	100	100	99	100	100	100	98	100	100	100
t_3	40	97	93	100	79	100	100	100	78	99	98	100

Accepted Manuscript

Table 4 *Test power for Scenario A.* Rejection frequencies (%) at asymptotic 5% level. Change-point tests based on variance (Var), mean deviation (MD), Gini's mean difference (GMD), and $Q_n^{0.8}$. **AR(1) process** with $\rho = 0.8$; scale changes by **factor** $\lambda = 2.0$.

Change location:	[*n*/4]				[*n*/2]				[3*n*/4]			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
n = 60												
$\mathcal{N}(0,1)$	13	12	25	27	19	13	32	34	15	8	22	21
$NM(3,0.01)$	12	11	24	27	19	14	32	32	15	8	21	20
$\mathcal{L}(0,1)$	10	10	19	25	17	11	29	31	11	7	19	18
t_5	10	8	20	22	17	12	30	30	13	8	20	20
t_3	9	8	18	21	14	9	24	27	10	6	17	17
n = 120												
$\mathcal{N}(0,1)$	9	10	26	27	26	22	48	41	16	10	28	17
$NM(3,0.01)$	10	10	24	24	26	22	46	40	15	10	25	15
$\mathcal{L}(0,1)$	8	9	20	22	20	17	40	35	16	11	17	24
t_5	8	9	22	22	20	18	41	34	15	12	24	17
t_3	6	5	15	16	13	15	28	25	10	8	19	14
n = 240												
$\mathcal{N}(0,1)$	13	19	39	36	52	58	75	62	45	38	57	34
$NM(3,0.01)$	12	19	36	35	47	54	71	60	38	32	49	30
$\mathcal{L}(0,1)$	10	28	14	26	41	49	65	54	30	27	43	26
t_5	10	14	26	25	38	48	66	52	32	31	47	28
t_3	6	11	20	18	25	36	49	40	21	23	36	21
n = 500												
$\mathcal{N}(0,1)$	42	66	79	75	95	98	99	97	90	89	94	86
$NM(3,0.01)$	37	61	74	70	91	98	98	98	86	88	92	86

Accepted Manuscript

Change location:	[$n/4$]				[$n/2$]				[$3n/4$]			
$L(0,1)$	30	55	68	63	88	95	96	94	80	81	87	76
t_5	28	55	65	61	84	95	96	94	76	80	86	76
t_3	14	38	44	41	58	80	82	81	53	65	70	59

Accepted Manuscript

Table 5 *Test size and power for Scenario B. (GARCH)* Rejection frequencies (%) at asymptotic 5% level. Change-point tests based on variance (Var), mean deviation (MD), Gini's mean difference (GMD), and $Q_n^{0.8}$.

Null hypothesis ($\lambda = 1$)												
Estimator:	Var	MD	GMD	$Q_n^{0.8}$								
n = 60	2	3	1	3								
n = 120	1	2	1	4								
n = 240	1	1	1	5								
n = 500	1	2	2	7								
Alternative ($\lambda = 2$)												
Change location:	[n/4]				[n/2]				[3n/4]			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
n = 60	2	4	6	6	9	13	19	18	9	10	14	12
n = 120	3	6	9	9	9	18	22	20	10	13	17	13
n = 240	4	12	13	13	16	36	36	34	19	29	30	26
n = 500	7	31	28	30	31	64	60	70	33	57	54	58

Accepted Manuscript

Table 6 *Test size for Scenario A.* Rejection frequencies (%) of change-point tests based on four different scale estimators at stationary sequences. 5% significance level based on dependent block **bootstrap** with 1000 bootstrap samples; sample sizes $n = 60, 120, 240$; three different innovation distributions. data-dependent blocklength selection.

	independence				$AR(1)$ with $\rho = 0.8$			
Estimator	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
n = 60								
$N(0,1)$	4	4	4	2	9	8	10	7
$L(0, 1)$	4	4	5	2	11	9	11	6
t_3	3	4	2	1	11	9	11	7
n = 120								
$N(0,1)$	4	4	4	4	8	7	9	7
$L(0,1)$	2	3	4	2	10	9	11	7
t_3	2	4	3	4	8	8	9	5
n = 240								
$N(0,1)$	4	4	4	5	5	6	6	7
$L(0, 1)$	3	4	4	3	6	6	7	7
t_3	2	4	3	2	6	7	8	5

Accepted Manuscript

Table 7 *Test power for Scenario A.* **Independence and AR(1)**; scale changes by **factor** $\lambda = 2.0$. Rejection frequencies (%) at 5% level based on dependent block **bootstrap** with 1000 bootstrap samples;

Change location:	[*n*/4]				[*n*/2]				[3*n*/4]			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
Independent observations												
n = 60												
N(0,1)	16	24	45	26	64	70	85	65	60	54	75	43
L(0,1)	6	10	21	8	23	38	54	23	27	32	48	18
t_3	4	7	17	6	17	35	48	17	18	25	40	12
n = 120												
N(0,1)	52	76	86	77	98	98	99	98	94	93	97	91
L(0,1)	17	41	49	24	67	88	91	63	60	74	78	49
t_3	8	30	34	14	35	74	76	46	38	61	65	34
n = 240												
N(0,1)	99	100	100	100	100	100	100	100	100	100	100	100
L(0,1)	50	89	90	74	96	100	100	98	93	98	98	92
t_3	20	74	69	50	58	96	94	86	58	90	88	80
AR(1) with $\rho = 0.8$												
n = 60												
N(0,1)	16	18	25	20	40	41	55	42	47	38	52	31
L(0, 1)	18	17	25	17	36	39	51	36	43	36	48	25
t_3	15	16	25	11	34	36	48	31	35	31	42	23
n = 120												
N(0,1)	22	28	41	34	66	67	80	63	61	52	66	48
L(0,1)	19	24	36	30	51	55	68	53	50	43	59	35

Accepted Manuscript

Change location:	[n/4]				[n/2]				[3n/4]			
t_3	19	21	32	22	41	49	63	45	40	37	49	34
n = 240												
$N(0,1)$	36	52	66	56	84	87	93	87	78	71	83	71
$L(0, 1)$	28	42	55	45	75	79	88	79	71	68	78	57
t_3	17	31	40	30	52	70	76	59	52	55	65	45

Accepted Manuscript

Table 8 *Test size and power for Scenario B. (GARCH)* Rejection frequencies (%) at 5% level based on dependent block **bootstrap** with 1000 bootstrap samples; scale changes by **factor** $\lambda = 2.0$.

Null hypothesis (λ = 1)												
Estimator:	Var	MD	GMD	$Q_n^{0.8}$								
n = 60	6	8	8	3								
n = 120	7	8	8	5								
n = 240	6	6	7	5								
Alternative (λ = 2)												
Change location:	[n/4]				[n/2]				[3n/4]			
Estimator:	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$	Var	MD	GMD	$Q_n^{0.8}$
n = 60	6	10	21	10	17	26	42	24	18	39	22	19
n = 120	10	17	22	15	24	41	50	33	26	36	46	28
n = 240	13	34	36	32	37	65	69	55	38	54	52	45

Accepted Manuscript